PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :		(11) International Publication Number:	WO 98/20647
H04L 12/18, 29/06	A1	(43) International Publication Date:	14 May 1998 (14.05.98)

(21) International Application Number: PCT/US97/20449

(22) International Filing Date: 7 November 1997 (07.11.97)

(30) Priority Data:

60/030,728	8 November 1996 (08.11.96)	US
08/937,129	24 September 1997 (24.09.97)	US

(71) Applicant: INTEGRATED TELECOM TECHNOLOGY [US/US]; Suite 300, 18310 Montgomery Village Avenue, Montgomery Village, MD 20879 (US).

(72) Inventors: CHAUDHRI, Imran; 11 Argosy Circle, North Potomac, MD 20878 (US). SEETHARAM, Srini, Wishnu; 9405 Royal Bennet Terrace, Gaithersburg, MD 20879 (US).

(74) Agents: LeBLANC, Stephen, J. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111 (US).

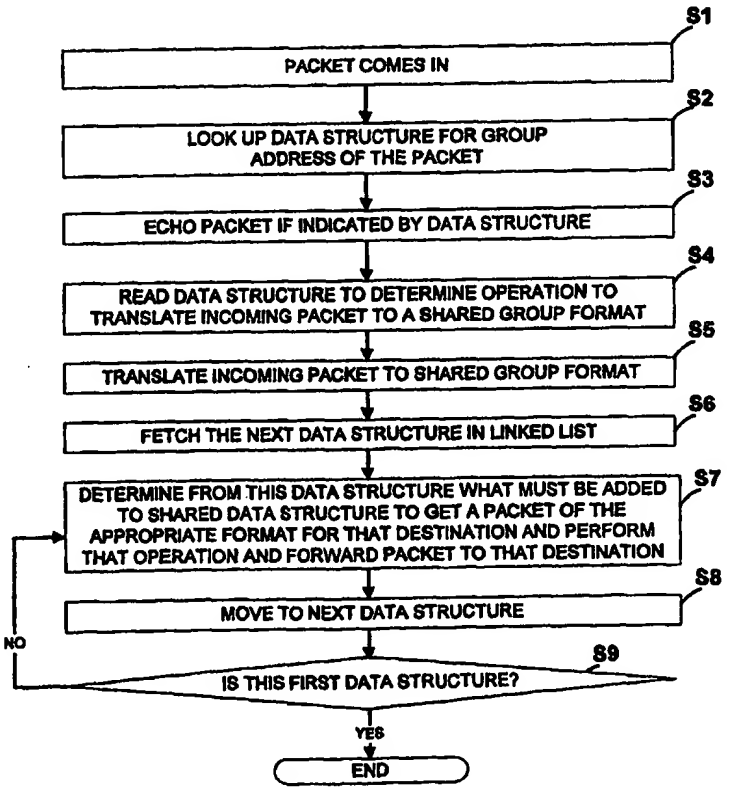
(81) Designated States: CA, JP, KR, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD AND APPARATUS TO TRANSLATE DATA STREAMS AMONG MULTIPLE PARTIES

(57) Abstract

Multiparty communications are handled in a digital device according to a data structure where a formatted packet from a particular transmitter is translated into a shared format M which is then retranslated to various destinations. Handling of group address packets can be accomplished with customized circuitry. Flexibility is achieved by allowing parameters of the routing routine to be specified in memory which can be programmably altered. In a specific embodiment, the representation of a group is a circular linked list. Each party has one entry in the linked list, therefore the entire structure has N substructures for N participants.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD AND APPARATUS TO TRANSLATE DATA STREAMS AMONG MULTIPLE PARTIES

This application claims priority from provisional patent application 60/030,728 filed 11/8/96, which is incorporated herein by reference. Related background information is discussed in related coassigned patent application serial number 08/937,129, filed September 15, 1997, entitled METHOD AND APPARATUS TO IDENTIFY FLOWS IN DATA STREAMS, which is incorporated herein by reference and which was filed with a microfiche appendix of 169 frames on two sheets comprising a preliminary user manual and programming information for an apparatus related to the invention. This appendix is also incorporated by reference.

A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

BACKGROUND OF THE INVENTION

The present invention relates to electronic circuits and to digital communication. More particularly, the present invention relates to an advanced method for handling data units (packets) within a communication device by efficiently translating data units to different formats among multiple parties using a simplified and efficient method.

The present invention has a number of applications in advanced telecommunication systems and networks. One envisioned application of the invention is in a state-of-the-art router (or layer three switch) for handling data packets in multiparty communications. These data packets may be constructed according to a number of well-known packet transmission protocols in a layered protocol suite. Routers are packet switching devices characterized by their ability to

perform intelligent routing of packets on a per-packet basis. Routers generally can handle packets or frames of data formatted according to a number of different protocols. A layer three switch is a type of router that performs layer three routing of packets largely or exclusively using dedicated logic circuitry.

Other applications of the invention include fast processing of any kind of discreet data unit streams quickly into multiple formats.

Explanation of both the prior art and the invention will be better understood with reference to specific examples, however this should not be taken to limit the invention to the particular examples described. The invention is generalizable to other similar types of communication devices or to other computing devices handling data in ways that are analogous to data handling in modern networks. The invention should therefore not be limited except as provided in the attached claims.

Networking Devices Standards

This specification presumes familiarity with the general concepts, protocols, and devices currently used in LAN networking and WAN internetworking applications as discussed in the above referenced applications.

Packets

In the most common LANs and WANs in use today, data is transmitted among network devices and processed within devices as a series of independent packets, each packet containing at least one header having at least a destination address specifying an ultimate destination and generally also having a source address. Headers may also include other handling information such as an indication of the different protocols used to format a packet, a transmission priority, etc.

Fig. 1 depicts an example of a packet as it may be transmitted on a network. The example shown is essentially an Ethernet packet, having an Ethernet header 202 and a 48-bit

Ethernet address (such as 00:85:8C:13:AA) 204, and an Ethernet trailer 230. Within the Ethernet packet 200 is contained, or *encapsulated*, an IP packet, represented by IP header 212, containing a 32 bit IP address 214 (such as 199.22.120.33). Packet 200 contains a data payload 220 which holds the data the user is interested in receiving or holds a control message used for configuring the network. Many other types and configurations of packets are presently known in the networking art and it is assumed that additional packet protocols will be developed in the future.

Multiple-Destination Data Units

A number of the networking standards referenced above provide a mechanism for a source in LAN or WAN to transmit a packet to all the nodes in the network. This situation may be referred to as a multipoint to multipoint communication.

A multipoint to multipoint communication is one where multiple calling parties are communicating with multiple called parties (e.g., a party line, or a conference call). In a packet-based communication system, a packet generated by any one party must be (possibly duplicated) and sent to all the other parties so that they can all participate in the conversation. This is known as a multicast operation.

In some cases, the party that generated the packet must also receive the packet it generated. This gives that party an indication that the packet it generated was multicast to all the parties in the communication. Within the same communication, it is also the case that certain parties may need this "echo" back to the generator of the packet, while other parties do not.

In a multiprotocol environment, each of the parties in the communication may be located on different types of networks. Where the eventual message content of each packet must be identical, the encapsulation of the packet is based on the network that is carrying it. For example, if the packet was generated on network type A, then it consists of a message section M and the encapsulation A (A:M). If the message must

be transported to two different networks B and C, then the packet on those networks will be in the format B:M and C:M, respectively. In addition, the original packet A:M may be sent back to the party on network A.

In prior art systems, translations between protocols are accomplished individually within a network device. For example, in a multipoint conversation between A, B, and C, each communicating over a different protocol, a prior art network device will typically store three data structures, each having two substructures. A data structure will be defined for each possible transmitter, A, B, and C, and a sub-structure will be defined for each translation, such as A->B, A->C, B->A, B->C, C->A, C->B. Adding a new party to the conversation, such as D, in such a prior art system will require adding a new data structure with three substructures, D->A, D->B, D->C, and also adding substructures to each of the existing transmitters A->D, B->D, C->D. In general, in these prior art systems, a multiparty conversation between N nodes requires N^2 data substructures for routing. This somewhat cumbersome system is necessary because prior art network devices generally handle routing all packets by first looking up a data structure for the transmitter of the packet and then processing the packet according to just that particular data structure.

What is needed is a mechanism that would allow a high performance communication device to quickly perform multipoint routing, possibly in hardware, while allowing the operation of the routing function and multicast group assignment to be specified and configured programmably so that it can be easily changed and can flexibly respond to different translation requirements of different network protocols.

SUMMARY OF THE INVENTION

In general terms, the present invention comprises techniques and devices for an improved handling of unitized or packetized data in a multipoint or similar transmission environment. More specific applications of the invention

include handling of multicast packet data in a network device such as a router or layer three switch.

According to the present invention, a multicast group G is defined. This group consists of all the parties that are participating in the communication. As parties join the communication, they are allowed to join G and each party can leave G when it desires. In a specific embodiment, the representation of G is a circular linked list. Each party has one entry in the linked list, therefore the entire structure has N substructures for N participants. This provides a number of advantages over prior art systems that typically require N^2 substructures for N participants. Each entry has five fields. The first field indicates the amount of encapsulation that must be removed from the incoming packet to extract the message M. Note that this is different for each party because each one exists on a different type of network. The second field indicates the amount and content of encapsulation that must be added to the message M in order to create a packet for that type of network. The third field indicates the amount and content of encapsulation that must be added to the message M in order to create a packet for that type of network. The third field is the "echo" bit that indicates if the packet must be sent back to the originator of the message. The fourth field is the pointer to the next entry of the group G. The fifth field indicates the destination address of the packet.

In a specific embodiment, the routine used to move through the circular linked-list and to translate data packets is performed by dedicated hardware logic, thereby allowing fast searching. However, the search is performed using a data structure stored in modifiable memory locations or registers. Thus, fast but flexible searching on packets may be performed.

The invention also comprises a network operating with devices having multipoint processing as described herein.

Additional aspects of the invention will be better understood upon reference to the following description of specific embodiments and the attached claims.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a diagram of a packet as an example of a type of data unit upon which the invention may be effectively employed.

FIG. 2 is a diagram illustrating a layered network protocol.

FIG. 3 is a flow chart illustrating a method of the invention.

FIG. 4 is a block diagram of a network device incorporating aspects of the invention.

FIG. 5 illustrates an example circular linked-list data structure within a memory according to the invention.

FIG. 6 is a block diagram of a search engine and memory according to a specific embodiment of the invention.

FIG. 7 is a block diagram of an alternative network device incorporating aspects of the invention

FIG. 8 is a diagram of a network according to one embodiment of the invention.

DETAILED DESCRIPTION OF SPECIFIC EMBODIMENTS

Multicast Routing Operation

FIG. 3 is a flow chart showing a multicast routing search operation according to one embodiment of the present invention and FIG. 4 is a block diagram of a network device (or element) 100 according to a specific embodiment of the invention.

When a packet 101 comes in on a port (Step S1), a multicast engine 104 looks up the data structure for the group address of the packet, indexed by the source address of the packet (Step S2). The data structure indicates whether the packet needs to be echoed, and the packet is eventually echoed if so indicated by the data structure (possibly without stripping the heading from the packet) either before, during, or after other forwarding of the packet (Step S3). The engine reads the data structure to determine what operation must be performed on the incoming packet to translate the packet to a shared group format M, which will generally be the message

portion of the packet (Step S4), and then performs that translation (Step S5). The engine looks in a pointer in the data structure to fetch the next data structure in the circular list (Step S6). This data structure specifies what must be added to M to get a packet of the appropriate format for that destination (Step S7). The engine performs the operation and forwards the packet (Step S7), and then moves to the next data structure (Step S8). Processing is repeated for each retrieved data structure until the data structure that the packet started with (Step S9) is reached.

One specific example of a data structure for storing information about each member of a multicast group is shown in FIG. 5. As shown in the figure, the example data structure includes a first translation parameter for removing the A header to get M, a second translation parameter for adding an A header to M, an echo bit, a pointer to a next data structure in the circular list, and a destination address for A. According to an embodiment of the invention, this single multicast group data structure is all the information storage that is required to implement multipoint to multipoint on multiprotocol networks. Using the example of three networks A, B, and C, the packet originates on network A. The multicast processor recognizes this and identifies the multicast group G. The processor remembers the origin of the packet was on network A. The processor then strips off the encapsulation of A and saves the shared format portion. The processor traverses down the linked list to the entry for B. It creates the encapsulation for B using the second field in the entry. It then proceeds down the linked list to the entry for C. The encapsulation for C is created using the second field in this entry. The processor traverses the link to the entry for A. The processor knows that this is the origin of the packet. If the echo bit is set, then the processor creates an encapsulation for A and it then stops traversing the linked list. If the echo bit is not set, the process stops traversing the list without creating a packet to be sent back to the origin on network A.

Once all the required packets (encapsulating plus message) are ready, they are sent to the destinations as indicated in the fifth field of the entry. This single group G handles packets originating from all parties participating in the multipoint to multipoint communication.

FIG. 6 shows a more detailed block diagram of routing engine 104 and memory 110 according to various specific embodiments of the invention. In this embodiment is depicted engine 104 including registers or similar memory locations for storing translation parameters such as header removal field 117a, header addition field 117b, echo flag 117c, pointer 117d, and destination address 117e. Engine 104 also includes a mechanism for storing an extracted message (M) 112 and a translated packet 116. According to one embodiment of the invention, these registers are the alterable parameters of engine 104 which will otherwise perform its function using fast dedicated custom logic circuits. These alterable registers may be set up to default values by a management program. According to the invention, engine 104 is able to very quickly translate and then route a data packet according to the values specified in the registers 117a-e.

While memory locations are shown within engine 104 for sake of illustration of an embodiment, it will be understood that these alterable values may be stored in any type of memory to which engine 104 has fast access, including in a portion of system memory 110.

Other Characteristics

The invention, in various embodiments, provides a number of benefits in high performance network devices or in computer systems handling similar data. According to the circular list of the invention, individual transmitters or receivers may be trivially added or dropped from the group. A new transmitter or receiver is simply added to the circular list by creating a single data structure for that transmitter or receiver, and then changing just one pointer in the existing circular list while copying the previously chained pointer into it's own pointer field.

According to another aspect of the invention, the format of the shared format M can be adjusted for different groups depending on whether the group members all have a common protocol or all reside on a subpart of the LAN, potentially simplifying processing. The invention also allows for distributed and parallel processing of multiple incoming packets by using multiple engines, each referring to the same linked list, an example of which is shown in FIG. 7.

It will also be apparent from the invention within the context of the prior art, that one implementation of the invention will allow for designing of dedicated fast logic circuitry to perform the translation of packets according to the data stored in the linked list, allowing the actual configuration of translation to be programmably configurable through setting data in the multipoint group data structure.

In one embodiment, the invention may be incorporated into a wide area network (WAN) or local area network (LAN) as shown in FIG. 8 to allow efficient handling of group traffic. FIG. 8 shows the invention as embodied in a network with a WAN backbone 300 and LANS 310a-c. LAN 310s may be connected to the WAN backbone by a local router 304.

The invention has now been explained in accordance with specific embodiments, however many variations will be obvious to those of skill in the art. In particular, the memory and data structure may be organized differently so long the general translation and linked-list method described herein may be performed. Also, the block diagram of FIG. 4 and the description organize and group functional blocks in a particular way to provide a clear example. However, other organizations are possible. These enumerated examples are not exhaustive of possible alternative embodiments of the invention. The invention should therefore not be limited except as provided in the attached claims.

BNSDOCID: <WO_____9820647A1_I_>

WHAT IS CLAIMED IS:

1. A method of routing data units in a digital device to multiple destinations comprising:

receiving a data unit from a source addressed to a group;

retrieving a data structure for said source to said group, said data structure containing translation instructions to a shared message component and a pointer to a next data structure in a linked list;

translating said data unit to a shared message component according to said translation instructions;

retrieving a second data structure according to said pointer, said data structure containing translation instructions from a shared message component and a pointer to a next data structure; and

translating said data unit from a shared message component according to said translation instructions.

2. The method according to claim 1 wherein said translating is performed by dedicated logic circuits.

3. The method according to claim 1 wherein parameters used in said translation are stored in memory and may be configured by operating system software of said digital device.

4. The method according to claim 1 further comprising repeatedly performing said second retrieving and translating steps until a pointer to the initial data structure is retrieved.

5. The method according to claim 1 wherein said translation may be performed by a plurality of translation engines in parallel such that a plurality of data units may be processed in parallel.

6. The method according to claim 1 wherein said translation is performed on data units consisting of encapsulated packets.

7. The method according to claim 1 wherein an incoming data unit is processed into a shared message component by:
removing header information from said data unit according to a rule stored in a register;
storing a resultant shared message component.

8. The method according to claim 1 wherein a substructure is created for each node added to a group of multiple destinations, such that the total number of data substructures for N nodes is of the order N.

9. A method of building a data structure for routing data units to multiple destinations comprising:
receiving a request from a node P to join a group G;
creating a substructure for said node P defining translation rules between said node P and a group message M; and
attaching said substructure to a linked list for said group G;
such that a group G consisting of N nodes comprises order N data substructures.

10. A method for processing packets having multiple destinations comprising:
setting up parameters in modifiable registers for translating incoming packets into a single shared format and for retranslating outgoing packets into different formats for different destinations; and
using dedicated logic circuits to apply said parameters to said packets to route packets to multiple destinations.

11. A network device for routing data units comprising:

an input for receiving a data unit directed to a group;

a search memory for storing a data structure for said group;

a search engine for performing a search for a group data structure; and

a packet processing routing unit for routing packets.

12. The device according to claim 11 wherein said data structure is stored in said memory as a linked list of entries, one entry for each destination belonging to a group.

13. A network for communicating data comprising:

a plurality of network devices capable of routing network data units between sources and destinations;

at least one network device comprising:

an input for receiving a data unit directed to a group;

a search engine for performing a search for a group data structure;

a search memory for storing a data structure for said group; and

a packet processing routing unit for routing packets.

14. A network according to claim 13 further comprising:

at least one network device capable of detecting a flow identification on an incoming data unit and routing said data unit according to said flow identification.

15. A network according to claim 14 further comprising:

at least one network device capable of adding a flow identification to an outgoing data unit.

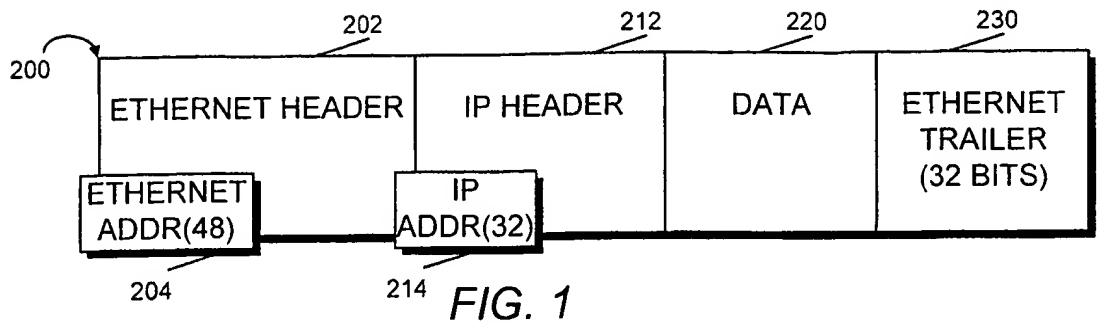


FIG. 1

	LAYER NAME (NUMBER)	DEVICES	DATA	PROTOCOLS
HIGH	HIGHER LAYER PROTOCOLS			
	APPLICATION LAYER (5)		FILES	FTP, HTTP
	TRANSPORT LAYER (4)	ROUTERS	ROUTING PACKETS	TCP, UDP
	ROUTING LAYER (3)	ROUTERS	ROUTING PACKETS	IP
	DATA LINK LAYER (2)	BRIDGES	PACKETS	ETHERNET
LOW	PHYSICAL LAYER (0,1)	REPEATERS	BITS	ETHERNET

FIG. 2

BNSDOCID: <WO_____9820647A1_I_>

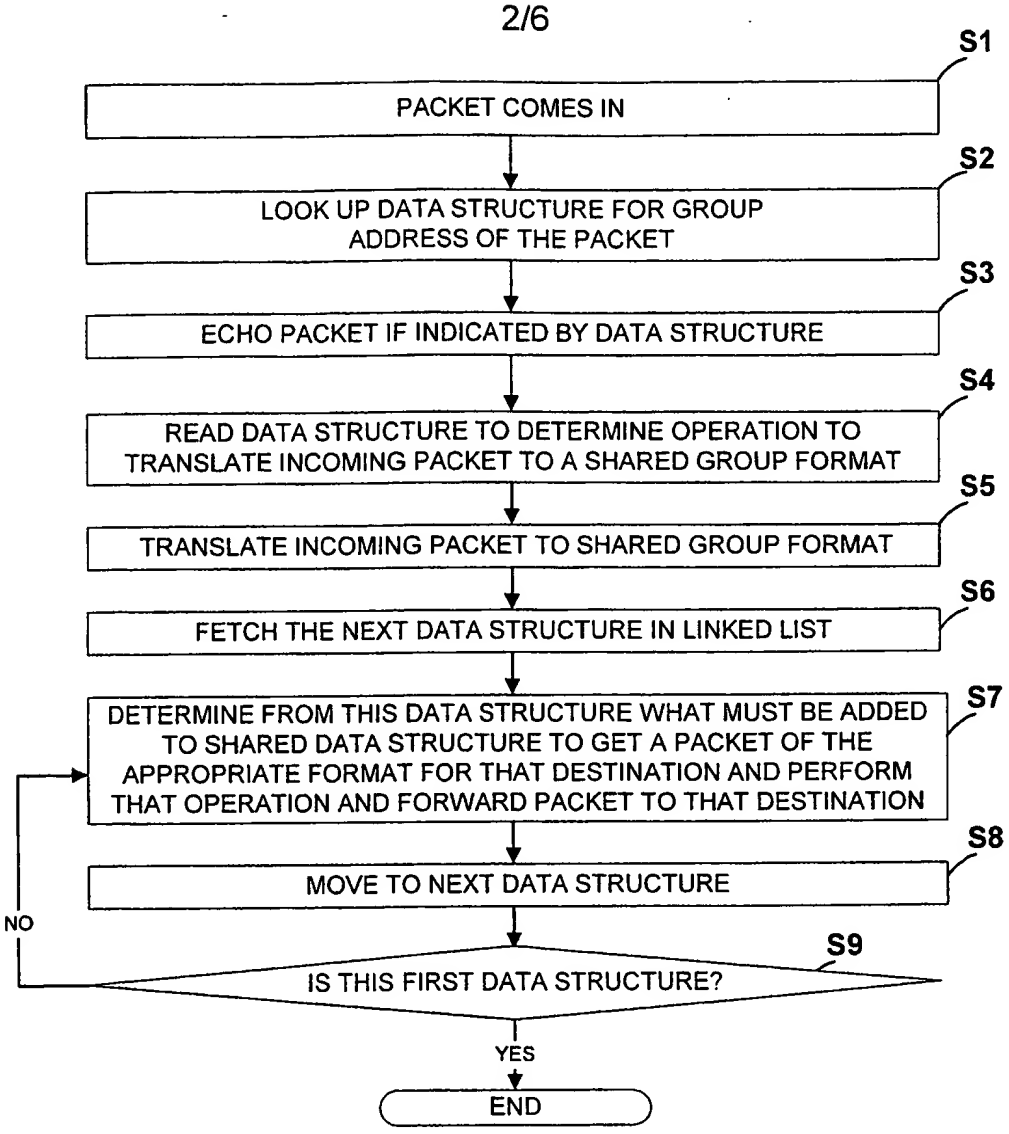


FIG. 3

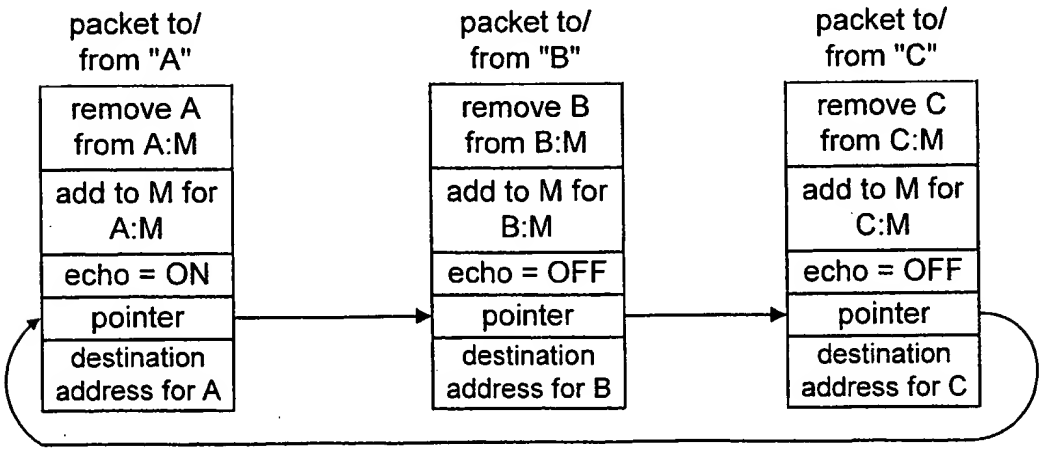


FIG. 5

BNSDOCID: <WO_____9820647A1_I_>

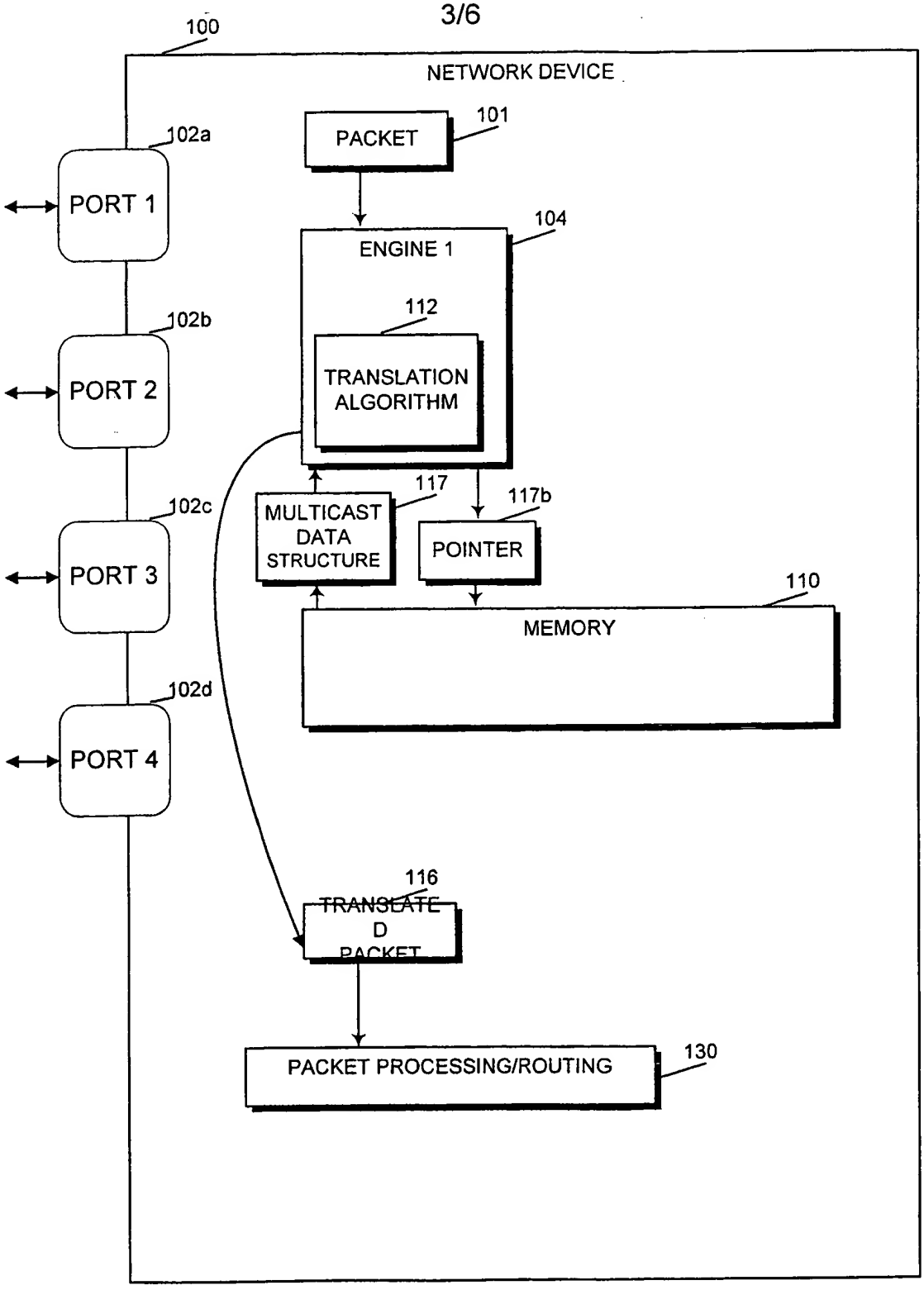


FIG. 4

BNSDOCID: <WO_____9820647A1_I_>

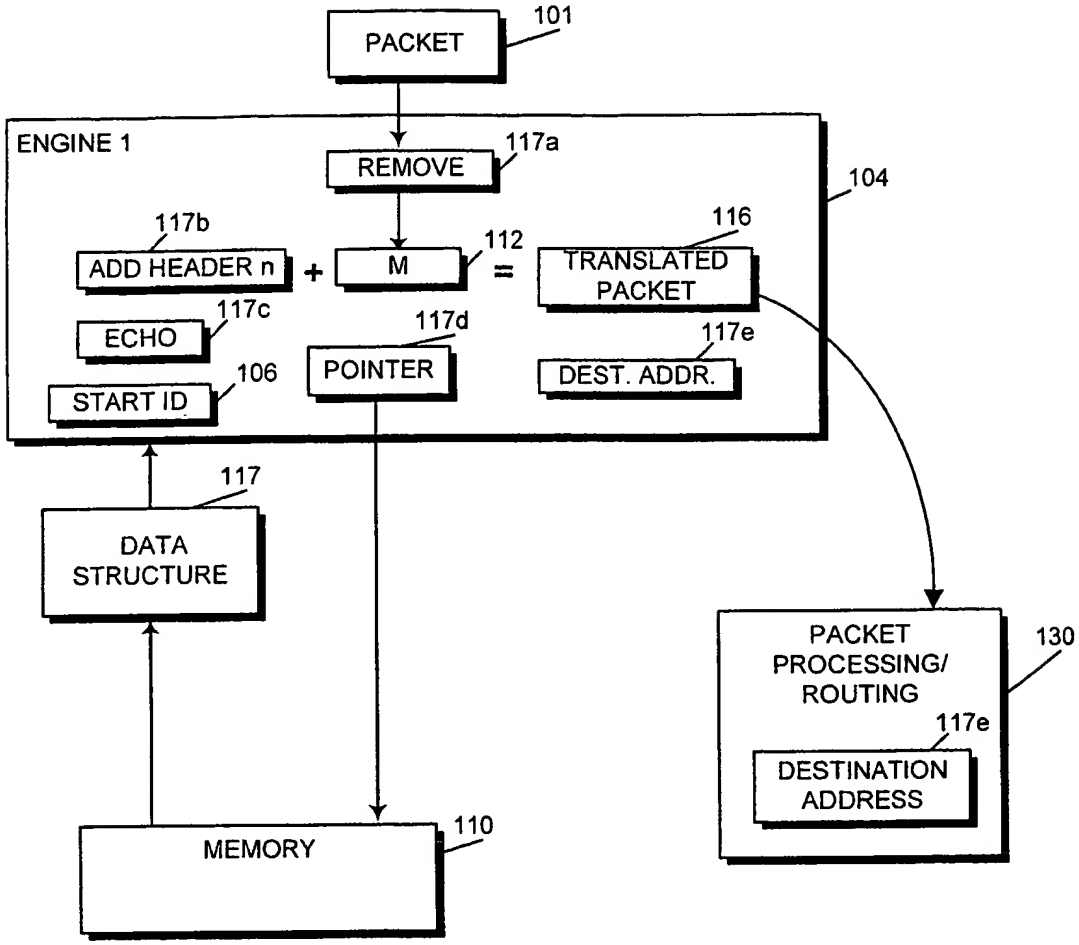


FIG. 6

BNSDOCID: <WO_____9820647A1_I_>

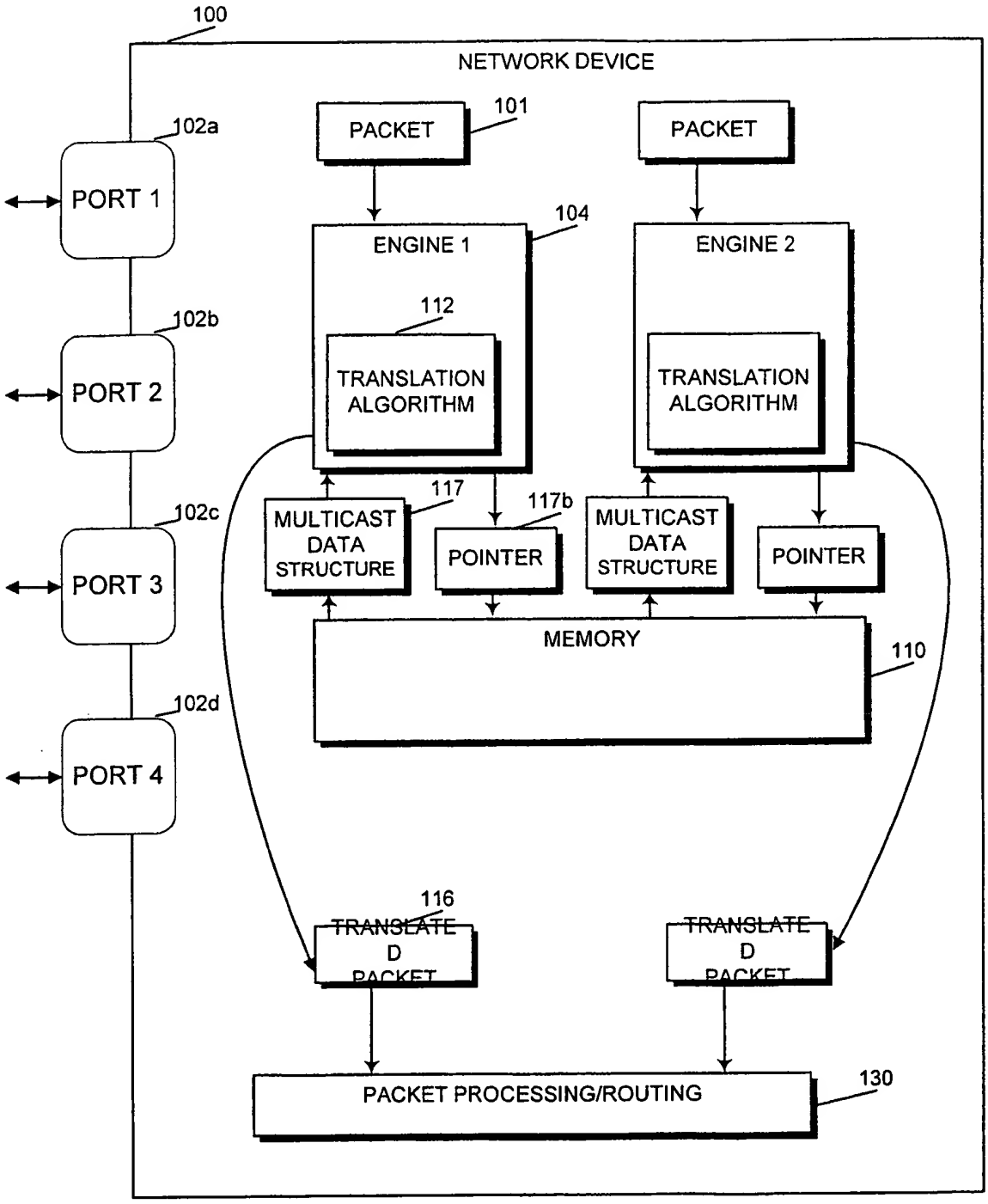


FIG. 7

BNSDOCID: <WO_____8820647A1_I_>

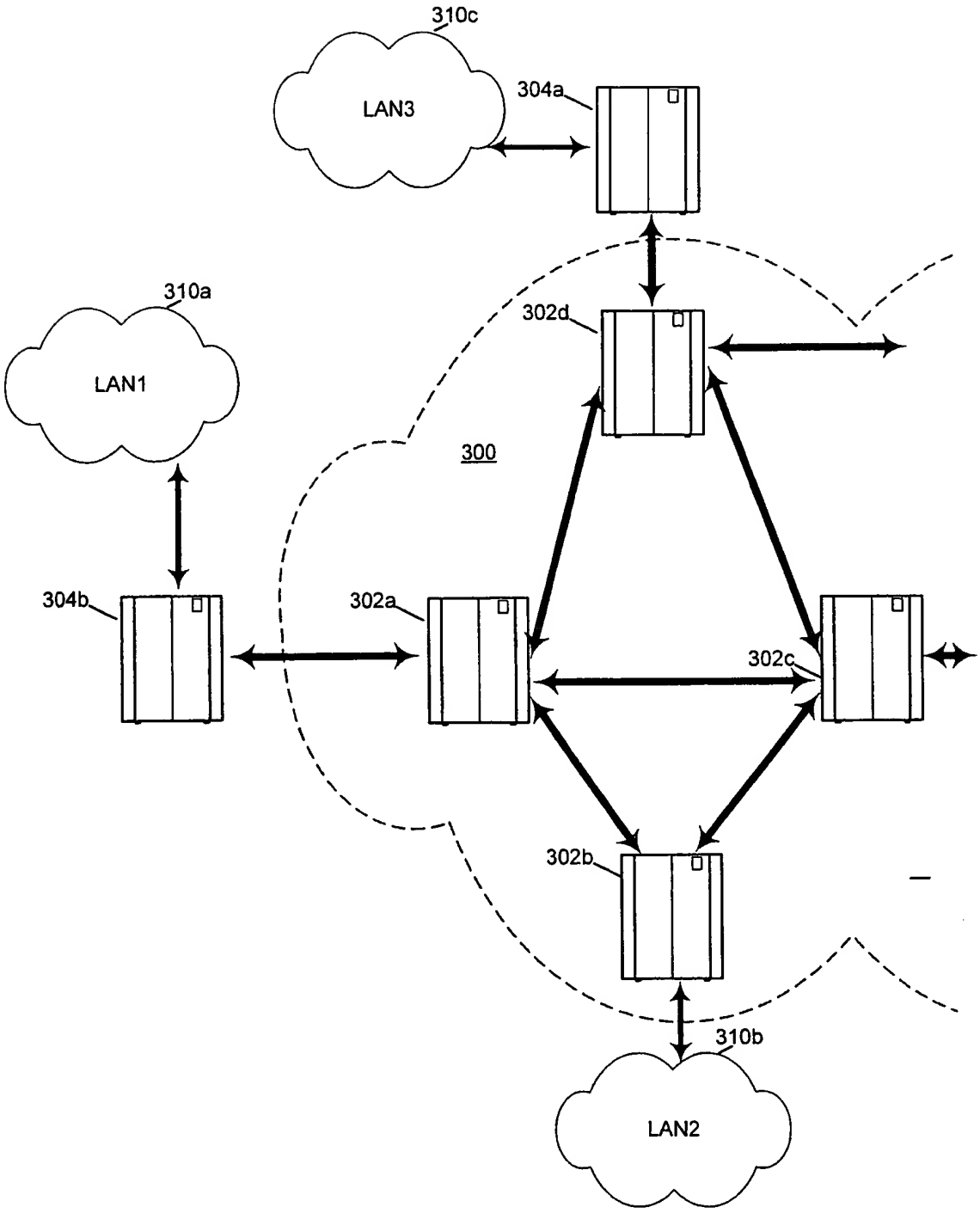


FIG. 8

BNSDOCID: <WO_____9820647A1_I_>

INTERNATIONAL SEARCH REPORT

Ir. ional Application No
PCT/US 97/20449

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 H04L12/18 H04L29/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	EP 0 666 670 A (IBM) 9 August 1995 see abstract see column 2, line 52 - column 3, line 3 see column 11, line 45 - column 12, line 32	11,13-15
A	see claim 1	1,7,9,10

Y	NGOH L H ET AL: "MULTICAST COMMUNICATION FACILITIES FOR DISTRIBUTED MULTIMEDIA INFORMATION SYSTEMS" SECOND IEE NATIONAL CONFERENCE ON TELECOMMUNICATIONS, 2 April 1989, UK, pages 271-276, XP000130145	11,13-15
A	see paragraph 3-5	8

	-/--	

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

1

Date of the actual completion of the international search	Date of mailing of the international search report
11 March 1998	25/03/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Perez Perez, J

Form PCT/ISA/210 (second sheet) (July 1992)

BNSDOCID: <WO_____9820647A1_I_>

INTERNATIONAL SEARCH REPORT

Ir ...tional Application No
PCT/US 97/20449

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication.where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 598 969 A (IBM) 1 June 1994 see claims	1,9-11, 13
A	"UNIVERSAL PROTOCOL CONVERSION" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 38, no. 12, 1 December 1995, page 323/324 XP000588157 see the whole document	1,6, 9-11,13

1

BNSDOCID: <WO_____9820647A1_I_>

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US 97/20449

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0666670 A	09-08-95	US 5491693 A	13-02-96
		CA 2130412 A	01-07-95
		JP 2571679 B	16-01-97
		JP 7212405 A	11-08-95
EP 0598969 A	01-06-94	BR 9304798 A	31-05-94
		CA 2105040 A	28-05-94
		CN 1091570 A	31-08-94
		JP 2539167 B	02-10-96
		JP 6224912 A	12-08-94
		KR 9614987 B	23-10-96
		US 5361256 A	01-11-94

BNSDOCID: <WO_____9820647A1_I_>